**Penalisation and shrinkage methods can produce unreliable clinical prediction models especially when sample size is small**

**Richard D. Riley1\*, Kym I.E. Snell1, Glen P. Martin2,**

**Rebecca Whittle1,Lucinda Archer1, Matthew Sperrin2, Gary S. Collins3,4**

\* corresponding author:

Professor of Biostatistics; e- mail: r.riley@keele.ac.uk; tel: +44 (0) 1782 733905

1 Centre for Prognosis Research, School of Medicine, Keele University, Staffordshire, UK. ST5 5BG

**2** Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK

3 Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK. OX3 7LD.

4 NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, OX3 9DU, United Kingdom

**Conflicts of interest:** None

**Funding**: Gary Collins is supported by Cancer Research UK (programme grant: C49297/A27294) and the NIHR Biomedical Research Centre, Oxford. Kym Snell is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR). This publication presents independent research funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. Glen Martin and Richard Riley are partially funded by the MRC-NIHR Methodology Research Programme [grant number: MR/T025085/1].

**Acknowledgements**: We thank two anonymous reviewers for constructive comments that helped us to improve the paper upon revision.

**Word count: 3900**

**Abstract**

***Objectives***

When developing a clinical prediction model, penalisation techniques are recommended to address overfitting, as they shrink predictor effect estimates towards the null and reduce mean-square prediction error in new individuals. However, shrinkage and penalty terms (‘tuning parameters’) are estimated with uncertainty from the development dataset. We examined the magnitude of this uncertainty and the subsequent impact on prediction model performance.

***Study design and setting***

Applied examples and a simulation study of the following methods: uniform shrinkage (estimated via a closed-form solution or bootstrapping), ridge regression, the lasso, and elastic net

***Results***

In a particular model development dataset, penalisation methods can be unreliable because tuning parameters are estimated with large uncertainty. This is of most concern when development datasets have a small effective sample size and the model’s Cox-Snell is low. The problem can lead to considerable miscalibration of model predictions in new individuals.

***Conclusions***

Penalisation methods are not a ‘carte blanche’; they do not guarantee a reliable prediction model is developed. They are more unreliable when needed most (i.e. when overfitting may be large). We recommend they are best applied with large effective sample sizes, as identified from recent sample size calculations that aim to minimise the potential for model overfitting and precisely estimate key parameters.

**Key words:** risk prediction models; penalisation; shrinkage; overfitting; sample size

1. **Introduction**

In healthcare, diagnosis and prognosis may be informed by statistical models that predict disease presence and outcome occurrence in individuals [1]. Such models are broadly known as clinical prediction models [2], and are often developed using a multivariable regression framework (e.g. logistic, survival, or linear regression), which provides an equation to estimate an individual's outcome probability (for binary or time-to-event outcomes) or outcome value (for continuous outcomes) conditional on values of multiple variables (predictors).

When estimating regression models using a particular dataset, conventional estimation techniques are often used, in particular ordinary least squares or standard maximum likelihood estimation. However, these tend to give model equations that are overfitted to the development dataset and so produce too extreme predictions when applied in new individuals; that is, some predicted outcome values lie too far from the overall mean. For example, using standard maximum likelihood estimation when developing a logistic regression model can give predicted probabilities too close to 0 for low-risk individuals, and too close to 1 for high-risk individuals. The problem of overfitting will usually increase as the sample size of the development data decreases, the number of candidate predictors increases, and (for binary or time-to-event outcomes) the number of outcome events decreases.

Penalisation and shrinkage methods have been proposed to resolve overfitting concerns. These include uniform shrinkage estimated via bootstrapping, ridge regression, the lasso (Least Absolute Shrinkage and Selection Operator), and elastic net [3-6]. Penalisation techniques shrink (in fact, introduce bias to) the estimated predictor effect estimates (i.e. odds ratios, hazard ratios or mean differences) towards the null. Compared to standard estimation methods, this reduces the variance of the developed model’s predictions in new individuals, thereby reducing the mean-square error of the predictions. For example, in a logistic regression model, penalisation methods will shrink predictor effects (odds ratios) towards 1, so that predicted probabilities in new individuals show less variability (i.e. are pulled away from 0 and 1, toward the mean outcome probability in the dataset).

Penalisation techniques are thus recommended as essential tools for prediction model development, [3, 7-12] especially for situations where the effective sample size is low (and thus potential magnitude of overfitting using standard methods is high) [13]. For example, Pavlou et al. conclude that “penalised regression is a flexible shrinkage approach that is effective when the EPV is low (<10)” [8], and Ambler et al. note that “the performance of ridge and lasso in our simulations suggests that it is possible, with care, to develop a risk model when the EPV is as low as 2.5” [7]. Given such recommendations, we are concerned that applied researchers might view penalisation methods as a ‘carte blanche’ to develop a prediction model regardless of the size of the dataset available for model development. Indeed, previous simulations suggest that although penalisation may indeed be effective on average, it may fail in the particular dataset being used for model development articles [12, 14].

In this article, we build on previous statistical articles [12, 14], to highlight this issue to a broad audience. We use applied examples and graphical displays to show that shrinkage and tuning parameters in penalised regression are typically estimated with large uncertainty. We demonstrate how this problem increases as the effective sample size reduces (i.e. when development datasets have smaller numbers of participants or events, relative to the number of candidate predictors) [15-17], and that the consequence is miscalibration and poor performance when the model is applied to new individuals. The impact on blood pressure predictions is demonstrated at the individual level, and our empirical examples are reinforced by analytic reasoning and a small simulation study. Section 2 describes our methods, Section 3 reports our results and applied examples, and Section 4 concludes with discussion and recommendations.

1. **Methods**

Penalisation methods for prediction model development are now described, followed by details of our applied examples and simulation study for evaluating their performance.

* 1. **Shrinkage and penalisation methods**

Perhaps the simplest penalisation method is where a uniform (linear and global) shrinkage factor () is used to shrink the predictor effects estimated from standard (unpenalised) maximum likelihood estimation. For example, a modified logistic regression model with shrunken predictor effects can be obtained by,

where is a shrinkage value between 0 and 1, and used to uniformly adjust the predictor effects ( estimated from standard maximum likelihood; and is the updated intercept, estimated after determining and applying to ensure that the calibration-in-the-large is correct (i.e. that the sum of predicted probabilities equals the overall proportion of observed events).

The true is that which minimises the expected mean-square error of predictions from the model when applied to the same population as that which the development dataset is sampled from. However, is an unknown parameter and so must be estimated, for example using the heuristic solution of Van Houwelingen and Le Cessie (given in appendix) [18], or via bootstrapping as described elsewhere [1, 2, 19, 20]. Riley et al. showed the estimate of depends on , which is the apparent (‘app’) value of the Cox-Snell (a measure of proportion of variance exampled by the model in the development dataset) [21].

Rather than using a post-estimation uniform shrinkage to penalise the regression coefficients from standard maximum likelihood estimates, other approaches are available that penalise during the model estimation itself [12]. In this article, we focus on three popular penalised regression methods: ridge regression [4, 6], the lasso [3], and elastic net [5]. Such penalisation approaches include a term , where is a particular penalty term and λ is a non-negative tuning parameter, which controls the amount of shrinkage. Further details of the penalty terms for each approach are given in the appendix. A value of λ=0 corresponds to no shrinkage (i.e. applying the standard maximum likelihood estimator when fitting a model such as logistic regression). The penalty factor for elastic net or the lasso is defined such that it can shrink predictor effects to zero and hence allows the exclusion of some predictors. The penalty term in ridge regression sets many of the predictor coefficients close to zero, but never exactly to zero.

As explained for the uniform shrinkage factor , the true value of the tuning parameter λ is the one that minimises the mean-square error of model predictions in the target population. However, λ is unknown and so is often estimated from the development dataset using either *K*-fold cross-validation, repeated *K*-fold cross-validation or bootstrap *K* -fold cross-validation [22]. Larger uncertainty in the value of λ leads to more uncertainty in the model’s actual predictive accuracy [23].

* 1. **Examples to illustrate uncertainty of uniform shrinkage estimate**

Three prediction models are used to illustrate uniform shrinkage, and the potential for uncertainty in the estimate of . Firstly, we developed two models for predicting systolic blood pressure (SBP) at 1 year in patients diagnosed with hypertension. The development dataset was based on a subset of Riley et al. [24], and to contrast models with different values, we developed models separately in those considered low risk (Model A, 262 participants with no co-morbidities) or high risk (Model B, 253 participants with co-morbidities) for developing cardiovascular disease. Seven predictors measured at baseline were used in the modelling: systolic blood pressure (mmHg), diastolic blood pressure (mmHg), body mass index (kg/m2), age (years), sex (female=0, male=1), current smoker (yes=1, no=0), and antihypertensive treatment (yes=1, no=0).

Secondly, we used data from 654 children to develop a model (Model C) to predict log-transformed forced expiratory volume (FEV in litres) using four predictors: age (years), height (inches), sex (female=0, male=1) and current smoker (yes=1, no=0). The data were obtained from <http://biostat.mc.vanderbilt.edu/DataSets>, and originally comes from Rosner [25].

* 1. **Simulation study to examine uncertainty of tuning parameter estimates**

Our simulation study to examine the degree of instability of the various penalisation methods is now described. The corresponding R code is provided at https://github.com/gscollins1973.

***Scenarios***

We considered 10 simulation scenarios. All scenarios had an outcome event proportion of 0.5 (50%) but varied in the sample size from N=100 to N=1000 (in steps of 100), corresponding to an events-per-parameter value of 2.5 (for N=100; 50 outcome events) to 25 (for N=1000; 500 outcome events). The scenarios are a pragmatic choice, to cover a range of events per parameter and to differ from those elsewhere [12, 14], but we recognise that they do not reflect all possible model development settings.

***Data generation***

For each scenario, 500 datasets of the chosen sample size were generated. Each dataset was simulated containing a binary outcome and values of 20 continuous predictors for each participant. Firstly, values of the twenty continuous predictors were simulated using a multivariate normal distribution with mean 0 and variance 1: five weakly correlated true predictors (pairwise correlation=0.1) and 15 uncorrelated noise predictors (pairwise correlation=0). Then, the true outcome () was generated conditional on an underlying logistic regression model based on all 20 predictors (equation given in appendix).

***Model development***

To each of the 500 simulated dataset in each scenario, six different methods were used to develop a logistic regression model based on the twenty predictors: three penalised regression methods (ridge regression, elastic net, and lasso), two uniform shrinkage methods (heuristic shrinkage or bootstrap shrinkage), and, for comparison, standard (unpenalised) regression. The tuning parameter (λ) for the ridge regression, elastic net and lasso was estimated using 5-fold cross-validation, using the cv.glmnet function from the glmnet package in R [26], to minimise the deviance. In addition to 5-fold cross-validation, we also investigated whether using bootstrapping would improve stability for ridge regression. In each bootstrap sample (total of 200 bootstrap samples), 5-fold cross-validation was used to select the tuning parameter, and the median λ over the 200 bootstrap samples was taken forward to develop the model.

***Model validation***

To evaluate the performance of each developed prediction model, a validation data set (N=5000; 2500 outcome events) was created using the same data generating procedure, giving a much larger effective sample size than the recommended 100 to 250 outcome events for validating a prediction model [27, 28]. Each developed model was evaluated in this independent validation data set by calculating the c-index, Nagelkerke R2, calibration-in-the large, and the calibration slope.

***Summarising simulation results***

For each type of model (e.g. uniform shrinkage, lasso, elastic net, ridge regression), in each scenario the key findings are the average model performance and, in particular, the variability in model performance across the 500 developed models in each scenario. Also, the variability in tuning parameter estimates. To illustrate this, boxplots are presented for each scenario.

1. **Results**

The key findings from our applied examples and simulation study are now described.

* 1. **Uncertainty in uniform shrinkage estimate: findings from applied examples**

Consider the three models to predict a continuous outcome introduced in Section 2.2. The corresponding unpenalised model equations are shown in Table 1, and cover situations with small (Model A), medium (Model B), and large (Model C) values for of 0.23, 0.56, and 0.81, respectively. For each model, Table 1 shows the mean bootstrap estimate of together with a 95% confidence interval for derived from the 2.5th and 97.5th percentile values of the bootstrap samples. Our ‘best guess’ of the true is the mean estimate across 1000 bootstrap samples but the width of the distribution of the 1000 values reveals the uncertainty in this choice. If the width is narrow, it gives more assurance that the mean estimate is suitable; however, if the width is wide, the mean estimate may be far from the true value of for the target population.

Table 1: Three prediction models developed using linear regression, with summary of model performance and bootstrap uniform shrinkage estimate

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Outcome** | **Model equation derived using ordinary least squares estimation (i.e. before any shrinkage)** | **Number of patients / predictor parameters** |  | **Uniform shrinkage () estimate from 1000 bootstrap samples**  **(95% confidence interval)** |
| **A** | Systolic blood pressure (SBP) (low CVD risk population) | 28.10 + 0.46\*SBP + 0.41\*DBP + 0.013\*BMI + 0.45\*age - 2.05\*sex - 17.81\*treat - 2.08\*smoker | 262 / 7  = 37 | 0.23 | 0.94  (0.77 to 1.18) |
| **B** | Systolic blood pressure (SBP) (high CVD risk population) | -12.69 + 0.94\*SBP + 0.21\*DBP -0.001\*BMI + 0.06\*age + 1.72\*sex - 1.04\*treat + 0.17\*smoker | 253 / 7  = 36 | 0.56 | 0.98  (0.87 to 1.10) |
| **C** | ln(FEV) | -2.07 + 0.02\*age + 0.04\*height + 0.03\*sex + 0.05\*smoker | 654 / 4  = 164 | 0.81 | 1.00  (0.96 to 1.04) |

For Model A, which has the smallest value (0.23) in the full development dataset and contains seven predictor parameters, the bootstrap mean estimate of is 0.94, which reflects low overfitting due to the larger number of 37 individuals per predictor parameter. Despite this, the range of observed bootstrap values for is reasonably wide, with 95% of values lying between 0.77 and 1.18. Thus we are not very confident about what the actual (true) shrinkage factor should be. Hence, even in an example where overfitting is expected to be low, the true shrinkage factor is very difficult to ascertain from the dataset at hand. Figure 1(a) shows that the uncertainty of S increases as the number of participants in the development dataset decreases, and as the corresponding mean estimate of moves closer to zero. For example, when refitting Model A using a random subset of 100 participants, the mean bootstrap estimate of is 0.82 and 95% of values fall between 0.52 and 1.27. When using a subset of only 50 participants, the mean estimate of is again smaller (0.78) and the 95% range of values wider (0.43 to 1.21). Thus, as the required level of shrinkage increases, the uncertainty in the shrinkage factor also increases.

**Figure 1** The mean estimate and 95% confidence interval of the uniform shrinkage factor ( as derived from 1000 bootstrap samples, across different sample sizes for developing models A, B and C as described in Table 1. Curves are created using a lowess smoother.

|  |
| --- |
| (a) is 0.23 for model A when using the maximum of 262 participants |
| (b) is 0.56 for model B when using the maximum of 253 participants |
| (c) is 0.81 for model C when using a maximum of 654 participants |

Now consider Model B, which was developed using 253 individuals and 7 predictors. These are similar numbers to Model A, but Model B has a much larger (0.56) in the full development dataset, a mean shrinkage estimate (0.98) closer to 1, and a narrower range of values for (95% interval: 0.87 to 1.10). To reduce sample size we sequentially removed one individual randomly at a time and repeated the modelling process; this shows again that the mean bootstrap estimate of reduces and the observed range of values widens (Figure 1(b)), especially as the sample size reduces below 100 individuals. For example, with only 50 individuals, the mean estimate of drops to 0.84 and 95% of bootstrap values lie between 0.52 and 1.10.

Model C has an of 0.81 in the full development dataset, much larger than those for Models A and B. The mean shrinkage estimate is 1 and the 95% interval is very narrow (Figure 1(c)); essentially, there is strong evidence that shrinkage of predictor effects is not required. This is not surprising as there are only 4 predictor parameters and 654 participants. Indeed, even if we developed the model in a reduced set of 20 randomly selected participants, the mean shrinkage estimate is still close to 1 and the 95% interval remains quite narrow (about 0.9 to 1.1).

Notice that, in these examples, there is less uncertainty in the value of when the estimate is closer to 1, i.e. settings where overfitting is less of a concern. Further, the estimate of is closer to 1 in settings where is closer to 1 (this is also shown analytically in the Appendix). Most prediction models do not have values close to 1. In particular, for binary and time-to-event outcomes the value of will often be much lower than 1, as the maximum value is bounded below 1 [16]. Consequently, uniform shrinkage estimates will often be far from 1 (towards 0) in models such as logistic and Cox regression, especially when the sample size is small and the number of predictor parameters is large.

**3.2 Importance of estimating shrinkage precisely: illustration using Model A**

To illustrate the importance of estimating precisely, we applied uniform shrinkage to revise the Model A equation shown in Table 1. Based on the full development dataset of 262 participants, bootstrap shrinkage suggests an of 0.94 with a 95% confidence interval between 0.77 to 1.18 (Table 1). We compared predictions after applying shrinkage values of 0.77 and 1.18. The difference in predictions was generally within 7-8mmHg for the majority of participants, and only 7 (2.7%) of the participants had a difference > 10mmHg (Figure 2(a)).

Next, we reduced the development dataset to a random sample of 50 participants, and re-estimating the equation for Model A. Due to the smaller sample size, the bootstrap approach gave a lower estimate of 0.78 for , with a wider 95% confidence interval of 0.43 to 1.21. We compared predictions after applying shrinkage values of 0.43 and 1.21. Compared to when using all 262 participants, the distribution of differences in predictions was much wider, and now 10 (25%) of the participants had absolute differences > 10mmHg; three participants even had a > 15mmHg absolute difference (Figure 2(b)). Hence, the larger the uncertainty in the shrinkage factor, the larger the uncertainty in predicted values for individuals, and so the greater the concern that model predictions may not be reliable for practice.

**Figure 2:** Difference in predicted systolic blood pressure (SBP) values in mmHg, when using the lower or upper bound of the bootstrap-derived 95% confidence interval for the shrinkage factor ( to revise Model A after (a) using 262 participants, and (b) using 50 participants for model development

|  |
| --- |
| (a) |
| (b) |

* 1. **Uncertainty in uniform shrinkage and penalised regression methods: findings from simulation study**

We now discuss the uncertainty when fitting penalised regression models, as identified by our simulation study of uniform shrinkage, ridge regression, the lasso and elastic net.

***Comparison of uncertainty in bootstrap and heuristic shrinkage estimates of***

For the uniform shrinkage approach, the simulations confirmed the findings identified from the applied examples. In particular, there was more variability in the estimate of uniform shrinkage with more overfitting (i.e. when was further from 1 towards 0), as shown in Supplementary material Figure S1. Additionally, the simulation results showed considerably less variability in the bootstrap estimate of than in the heuristic shrinkage estimate of when using small sample sizes for model development and when the required shrinkage is large ( (Supplementary material Figure S1). When the sample sizes were larger such that required shrinkage is smaller (), the bootstrap and heuristic approaches were much more similar in their distribution of estimates. This again suggests that estimates of are more stable when the overfitting is small (i.e. is closer to 1).

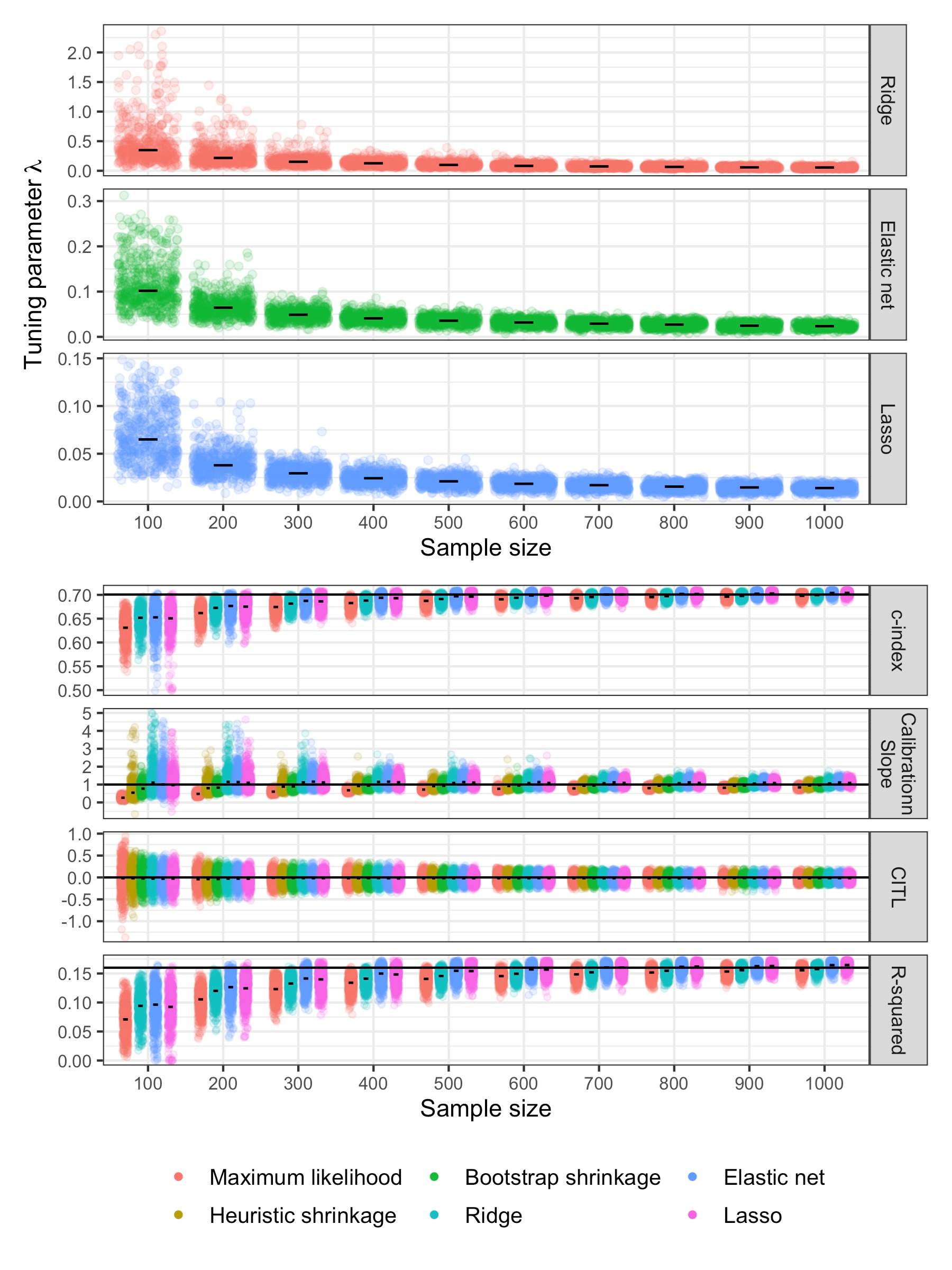
***Uncertainty in tuning parameter estimates and prediction model performance***

Figure 3(a) illustrates the variability in the estimation of the tuning parameter, , in our simulation scenarios for each of the three penalised regression approaches. For all three approaches, the smaller the sample size, the more uncertainty in the estimate of . Variability in the estimate of the λ was reduced using bootstrap 5-fold cross-validation, compared to standard 5-fold cross-validation for ridge regression (Supplementary Figure S5). There was negligible difference between 5-fold and 10-fold cross-validation (supplementary Tables S1 to S3).

Importantly, Figure 3(b) shows that more variability in led to more variability in the predictive performance of the developed models when tested in the large validation dataset (also see Supplementary Material Figures S2, S3 and S5). Even with the best model development approaches, there was considerable variability in a model’s predictive performance upon validation, especially when the development dataset had a small sample size. In particular, for every model there was large variability in the calibration slope upon validation, and the slope was often very different to 1 (i.e. indicative of miscalibration), with some methods even giving a range of about 0 to 5 in the smallest sample sizes (Figure 3(b)). Therefore, despite their intention, the various penalisation and shrinkage methods did not provide a ‘carte blanche’ that ensured a reliable model when applied in new individuals.

Large development sample sizes were needed to reduce the variability in both the estimate of and predictive performance upon validation, in order for the penalisation and shrinkage methods to give a more reliable model for practice (i.e. as the model development dataset became larger, the methods produced models with calibration slopes more consistently close to 1 when applied in new data). In that situation, the various penalisation and shrinkage methods performed similarly (Figure 3(b), Supplementary material Figure S5), with an average calibration slope estimate close to 1 and reasonably narrow variability (e.g. 0.8 to 1.2). Other measures were also more stable, such as , the c*-*index and the calibration-in-the-large (Figure 3(b)), and we would expect this to also hold for other measures such as the integrated or estimated calibration index [29, 30].

**Figure 3:** Median values (short horizontal lines) and scatter plots showing variability in (a) tuning parameter estimate (), and (b) predictive performance of developed model in the large validation data, for various methods across varying sample sizes for model development; for each sample size, 500 datasets were simulated as described in Section 2.3 and for each dataset a model was developed for each method with (a) tuning parameter estimated, and then (b) predictive performance tested. In (b) the long horizontal lines are the large sample performance values.



(a)

(b)

CITL = calibration-in-the-large; Horizontal spread within each sample size grouping is just random jitter to aid display; c-index is not shown for heuristic or bootstrap shrinkage, as these methods do not change the ranking of predictions, and thus the c-index is the same as maximum likelihood estimation.

1. **Discussion**

Our findings emphasise the potential for large uncertainty of shrinkage and tuning parameter estimates used within penalisation methods when developing prediction models. Though penalisation methods are recommended because they will improve upon standard estimation methods *on average*, in a particular dataset they can be unreliable. The most problematic datasets are those with small effective sample sizes and where the developed model has an far from 1, which is common for prediction models of binary and time-to-event outcomes [16]. Such uncertainty might lead to considerable miscalibration of predictions when the model is applied to individuals outside the development dataset.

A limitation of our work is that the simulation study scenarios were pragmatic, and so do not cover every possible type of model development dataset and setting. However, the findings echo related simulation studies by Van Houwelingen [12] and Van Calster et al [14]. Further, our simulation scenarios differ from those used in these papers, and we also examined the reduction in uncertainty in estimating tuning parameters over k-fold cross-validation using repeated k-fold cross-validation and bootstrap k-cross-validation, and assessed how this impacts on reducing uncertainty in model performance measures. We also demonstrated the issues using analytic reasoning and applied examples, to showcase the problem to a wider audience.

***Recommendations***

In model development datasets with large potential for overfitting, the uncertainty about the true magnitude of penalisation will often be very large, and thus there is actually no guarantee they will improve calibration of predictions in new individuals. Indeed, in situations where penalisation methods are most needed, they are more likely to unreliable. Hence, we recommend that researchers are best applying penalisation methods to develop a prediction model when the effective sample size is large, such that the amount of shrinkage and penalisation is anticipated to be small. In recent guidance [15-17], we show how researchers can base sample size requirements on a targeted uniform shrinkage factor of at least 0.9, such that the magnitude of global shrinkage is desired to be 10% or less. In such situations, penalisation methods will have narrower uncertainty about the estimated tuning and shrinkage parameters, such that developed models are more likely to be reliable (well calibrated) when applied to new individuals. This idea is supported by Figure 1, which shows that the range of observed bootstrap values for the shrinkage factor () become more acceptably narrow for sample sizes where the mean estimate of is between 0.9 and 1.0. Conversely, where the estimate of is less than 0.9 the variability in bootstrap values may escalate quite rapidly. Even when is estimated close to 0.9 or above, the uncertainty may be surprisingly large still in some applications (e.g. see confidence interval for for Model A, Table 1). Sample size should also be large enough to estimate overall model fit precisely and key parameters such as the model intercept [15-17].

Another key finding is that in smaller samples, bootstrapping resulted in less variability in estimates of shrinkage than other methods, including the heuristic estimate of uniform shrinkage and k-fold cross-validation estimate of tuning parameters in penalised regression. However, in situations where sample size is larger, the various methods were similar in terms of variability of shrinkage (and subsequent predictive performance upon validation). Hence, when the sample size is adequate, we recommend that any of the studied penalisation or shrinkage methods can be used, as they should perform similarly, and better than unpenalised regression unless sample size is extremely large and is large (e.g. as in Model C).

***Summary***

In summary, penalisation and shrinkage methods should not be viewed as a solution to small effective samples sizes for prediction model development. We recommend they are best applied to develop models in situations where a sufficiently large development dataset is available, to minimise the potential for model overfitting, improve the precision of model parameter estimates (including shrinkage and tuning parameters), and thus give more robust prediction models for clinical practice.

**What this study adds:**

* When developing a clinical prediction model, penalisation techniques are recommended to address overfitting, however they are not a ‘carte blanche’.
* Though penalisation methods will, on average, improve upon standard estimation methods, in a particular dataset they can be unreliable, as their unknown shrinkage and tuning parameter estimates are often estimated with large uncertainty.
* The most problematic datasets are those with small effective sample sizes and where the developed model has a Cox-Snell far from 1, which is common for prediction models of binary and time-to-event outcomes.
* Penalisation methods are best used in situations when a sufficiently large development dataset is available, as identified from sample size calculations to minimise the potential for model overfitting and precisely estimate key parameters.
* When the sample size is adequately large, any of the studied penalisation or shrinkage methods can be used, as they should perform similarly, and better than unpenalised regression unless sample size is extremely large and is large .

**APPENDIX:**

**Further description on shrinkage and penalisation methods**

The heuristic shrinkage factor of Van Houwelingen and Le Cessie is defined as,

*Equation (A1)*

where, is the total number of predictor parameters for the full set of candidate predictors (all those considered for inclusion in the model) and is the likelihood ratio (chi-squared) statistic for the model. Riley et al. show that the heuristic shrinkage factor can be re-written as [16],

*Equation (A2)*

where is the apparent (‘app’) value of the Cox-Snell (a measure of proportion of variance exampled in the model development dataset) [21].

The log-likelihood of penalised regression approaches can be expressed generally in the form , where is the penalty term and λ is a non-negative tuning parameter, which controls the amount of shrinkage. The actual penalty term varies based on the penalised approach. For example, the penalised log-likelihood for the elastic net takes the form:

Here, is the log-likelihood of the model (e.g. logistic regression) without penalisation,  is the mixing parameter and ranges between 0 and 1. An  is equivalent to ridge regression, and an  is equivalent to lasso. Whilst we could simultaneously tune over both and λ, for simplicity in this paper we chose for the elastic net applications.

**Further description of simulation study set-up**

After generating individuals’ values of twenty predictors ( to ), the true outcome ( = 0 or 1) was generated for each individual based an underlying logistic regression model of

, where the linear predictor () was

)

and the intercept ( was set to zero. The probability of an outcome event was calculated for each individual using the linear predictor as , and the true outcome (0 or 1) generated as a Bernoulli variable.

**Shrinkage is closer to 1 when R-squared is closer to 1**

Equation (A2) reveals that, for a particular number of participants ( and predictor parameters (, the heuristic shrinkage factor ( becomes closer to zero as the value of moves closer to zero. These analytical results are illustrated in the figure below, for a hypothetical prediction model of a continuous outcome developed using 5 predictor parameters; the estimate of decreases exponentially toward 0 as the assumed moves from 1 to 0. For small reductions in , the rate of change in the required value is relatively steep for values of < 0.8, and relatively flat when > 0.9. This indicates that there is less uncertainty in the estimated value of when the true is closer to 1, i.e. settings where overfitting is less of a concern.



References

[1] Riley RD, van der Windt D, Croft P, Moons KGM. Prognosis Research in Healthcare: Concepts, Methods and Impact. Oxford, UK: Oxford University Press; 2019.

[2] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.

[3] Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Statist Soc B. 1996;58:267-88.

[4] Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics,. 1970;12:55-67.

[5] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2005;67:301-20.

[6] Hoerl AE, Kennard RW. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics. 1970;12:69-82.

[7] Ambler G, Seaman S, Omar RZ. An evaluation of penalised survival methods for developing prognostic models with rare events. Stat Med. 2012;31:1150-61.

[8] Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ. 2015;351:h3868.

[9] Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of Shrinkage Techniques in Logistic Regression Analysis: A Case Study. Statistica Neerlandica. 2001;55:76-88.

[10] Copas JB. Regression, Prediction and Shrinkage. Journal of the Royal Statistical Society Series B (Methodological). 1983;45:311-54.

[11] Copas JB. Using regression models for prediction: shrinkage and regression to the mean. Stat Methods Med Res. 1997;6:167-83.

[12] Van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. Statistica Neerlandica. 2001;55:17-34.

[13] Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001;21:45-56.

[14] Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Stat Methods Med Res. 2020:962280220921415.

[15] Riley RD, Ensor J, Snell KIE, Harrell FE, Jr., Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.

[16] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Jr., Moons KG, et al. Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcomes. Stat Med. 2019;38:1276-96.

[17] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Jr., Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. Stat Med. 2019;38:1262-75.

[18] Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med. 1990;9:1303-25.

[19] Harrell FE, Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis (Second Edition). New York: Springer; 2015.

[20] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1-73.

[21] Cox DR, Snell EJ. The Analysis of Binary Data (second edition). London: Chapman and Hall; 1989.

[22] Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics. 2005;21:1979-86.

[23] Roberts S, Nowak G. Stabilizing the lasso against cross-validation variability. Computational Statistics & Data Analysis. 2014;70:198-211.

[24] Riley RD, Kauser I, Bland M, Thijs L, Staessen JA, Wang J, et al. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. Stat Med. 2013;32:2747-66.

[25] Rosner B. Fundamentals of Biostatistics (5th ed). CA: Duxbury: Pacific Grove; 1999.

[26] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33:1-22.

[27] Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35:214-26.

[28] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016; 74:167-76.

[29] Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine.0.

[30] Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B. Assessing calibration of multinomial risk prediction models. Stat Med. 2014;33:2585-96.

**Additional Figures and Tables**

Figure S1: Illustrating the variability in heuristic and bootstrap shrinkage factor by varying sample size

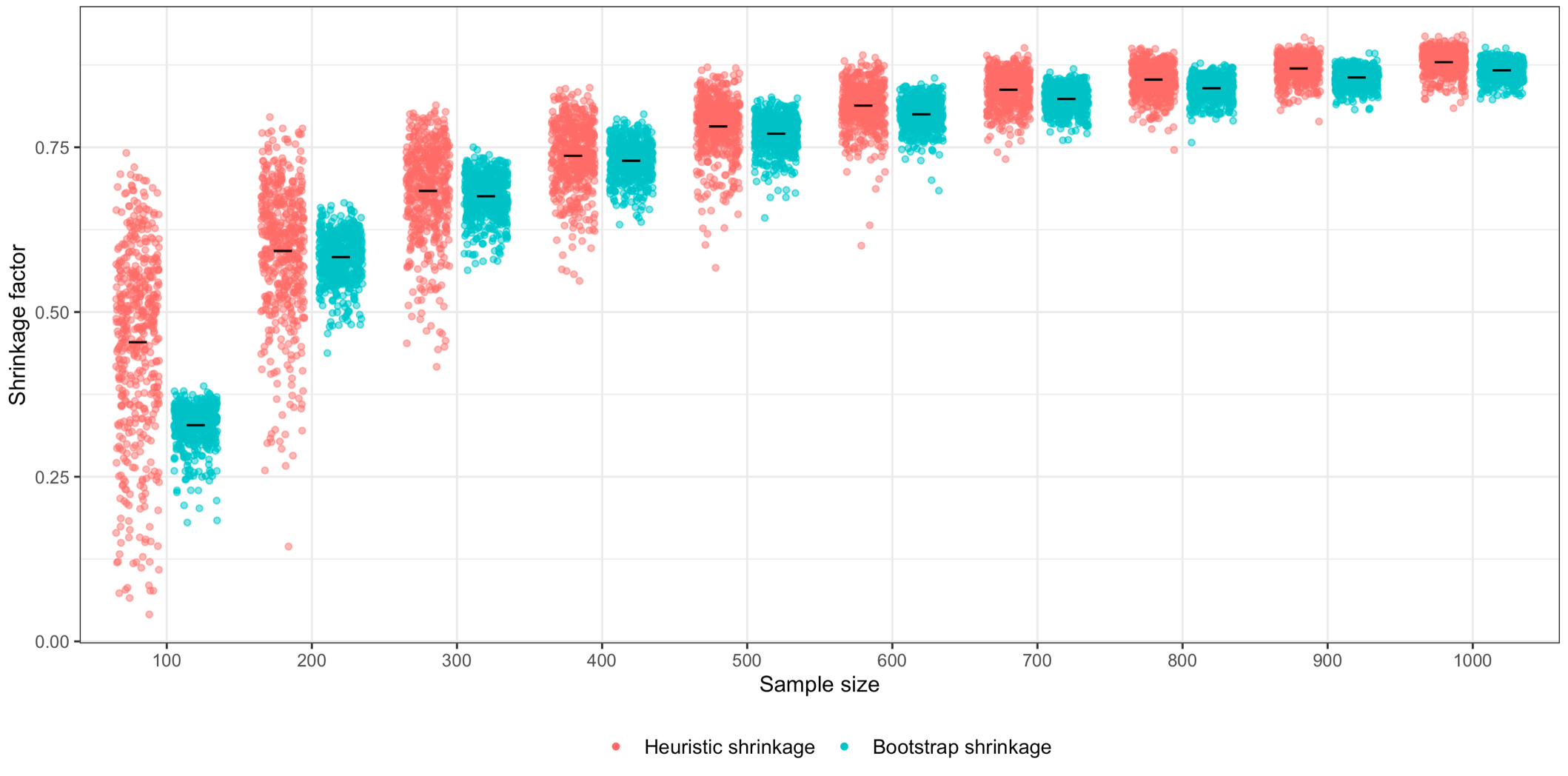
**

Figure S2: Apparent and validation c-index for models developed with no shrinkage (maximum likelihood) and penalised regression (ridge, elastic net, lasso)

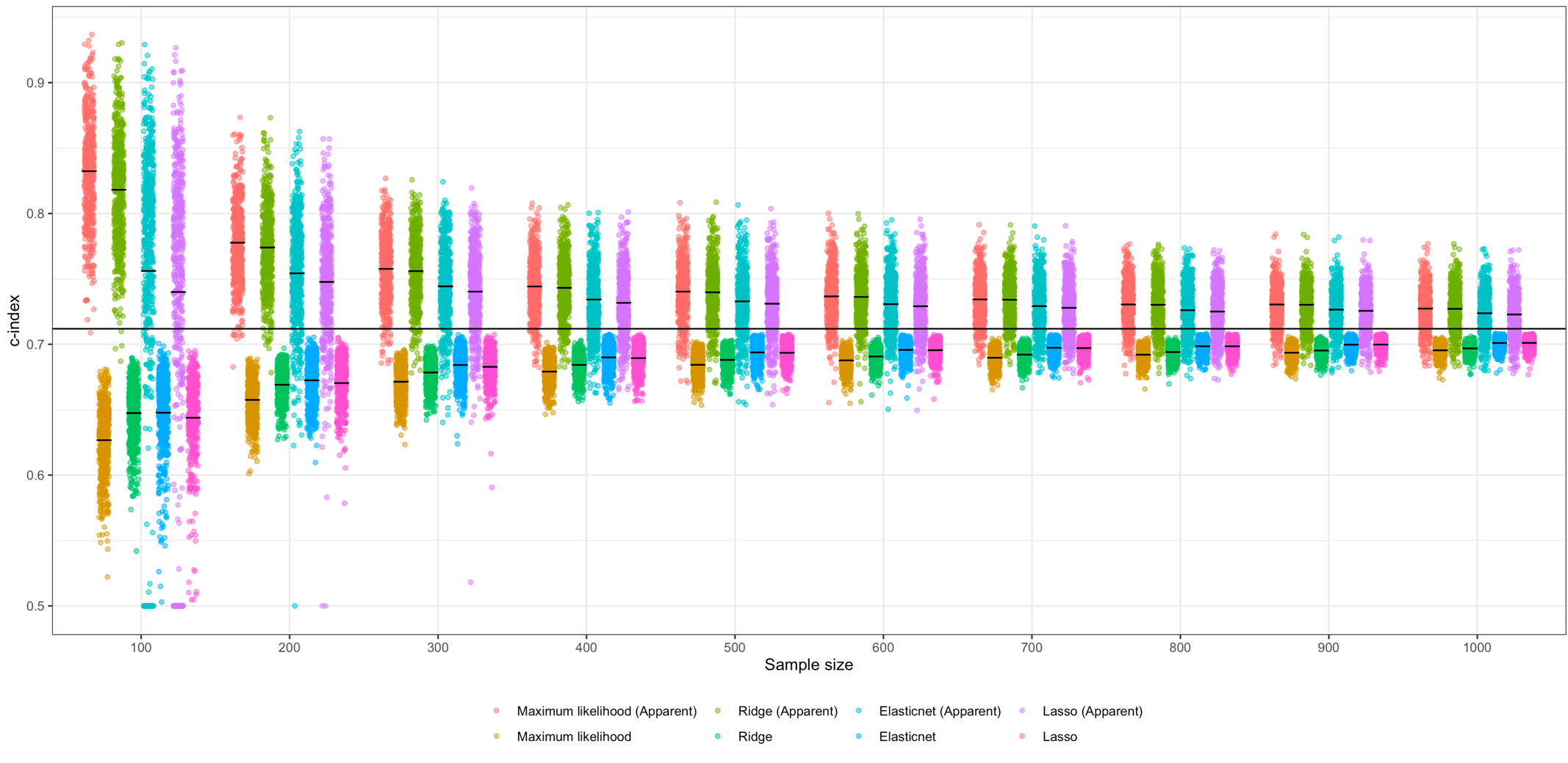


Figure S3: Calibration slope as estimated in the validation sample for models developed with no shrinkage (maximum likelihood), shrinkage (heuristic and bootstrap), and penalised regression (ridge, elastic net and lasso)

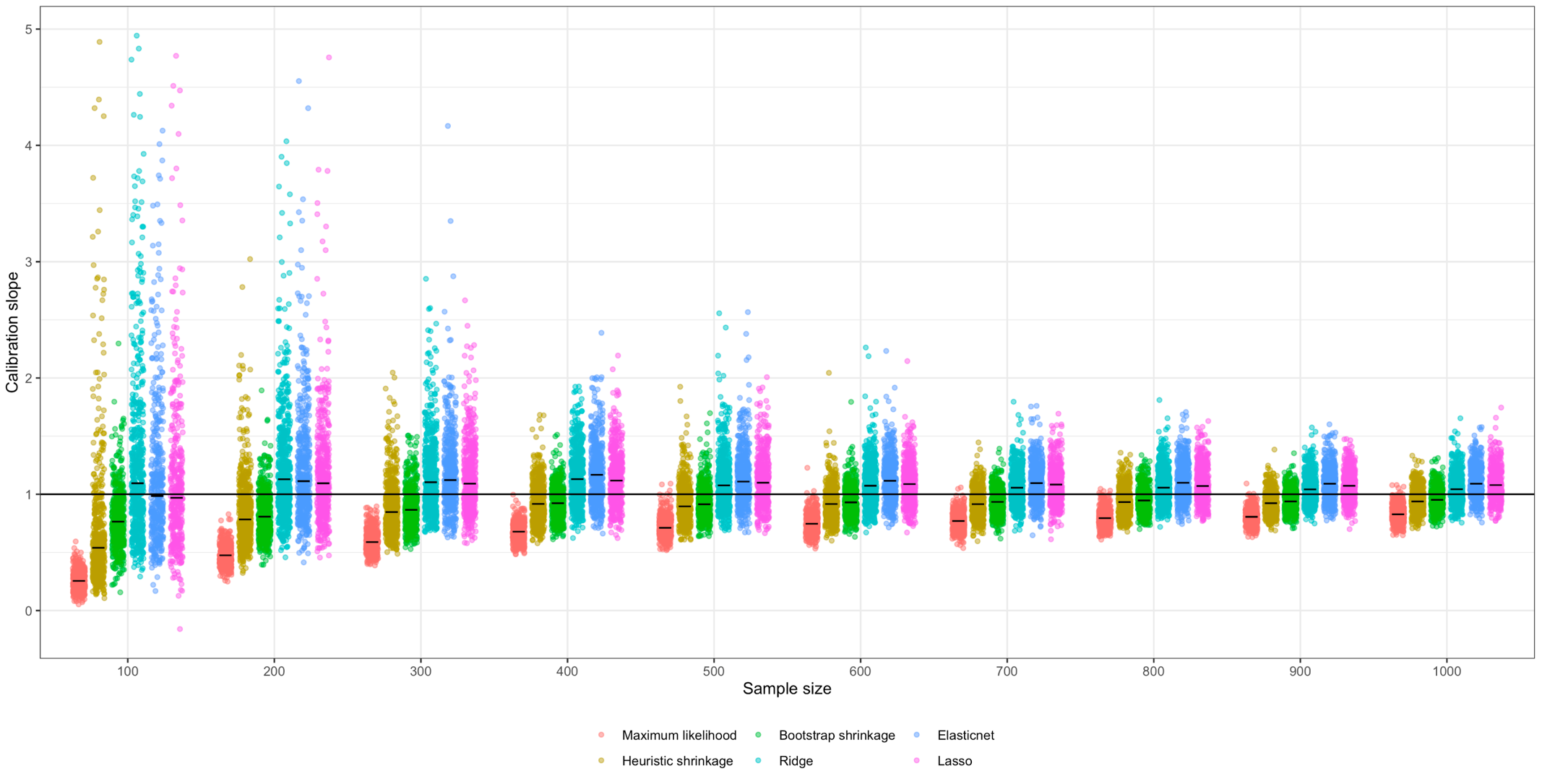


Figure S4: Standard deviation of the linear predictor for models developed with no shrinkage (maximum likelihood), and ridge regression (using 5-fold cross-validation, and bootstrap 5-fold cross-validation)

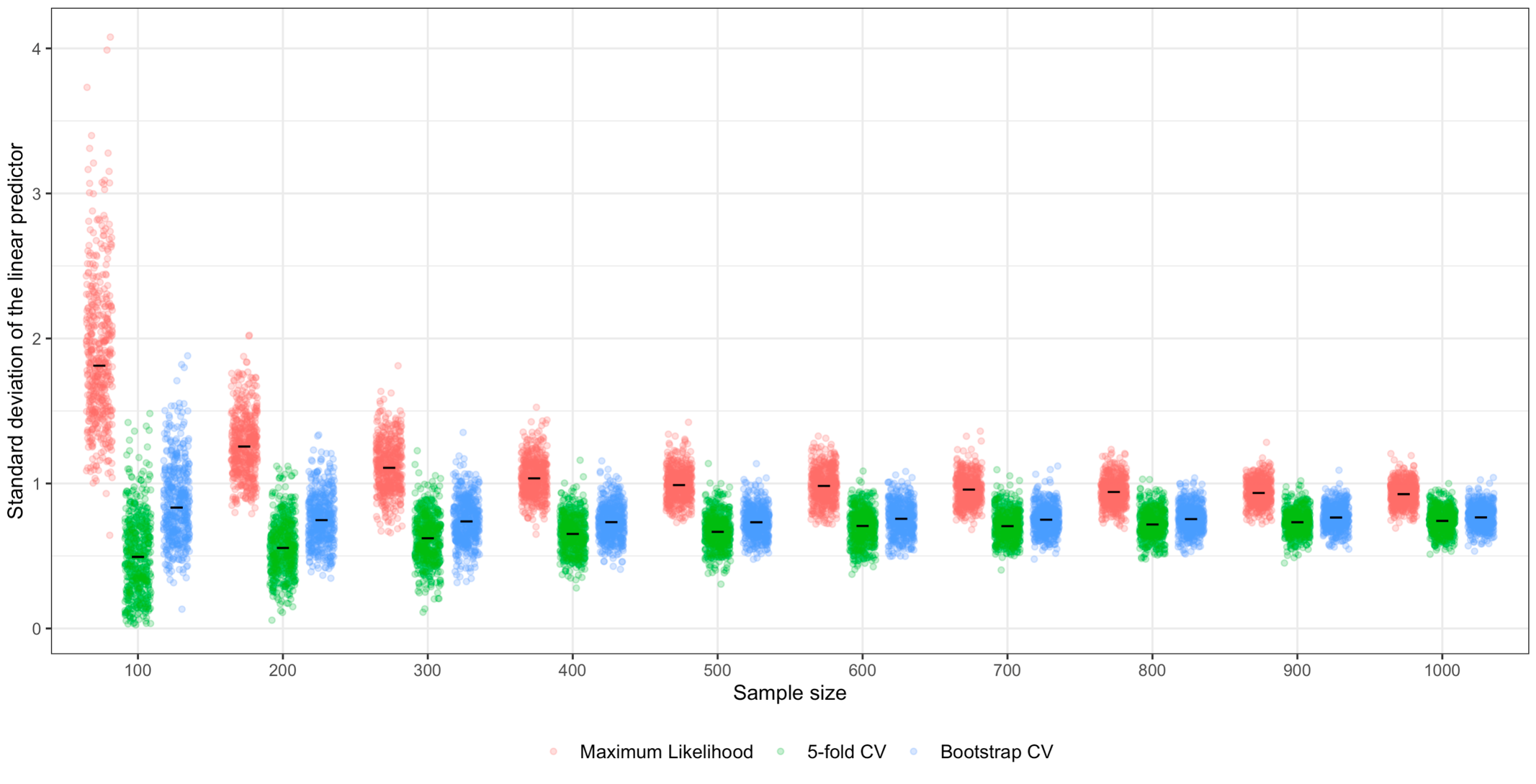


Figure S5: Box-plots and scatter plots showing variability in tuning parameter estimate from model development using ridge regression (panel (a)), and the estimated c-index (panel (b)) and calibration slope (panel (c)) of the developed model when tested in the large validation data, across varying sample sizes for model development. For each sample size, 500 datasets were simulated as described in Section 2.3 and for each dataset a model was developed using ridge regression with either 5-fold cross-validation or bootstrap 5-fold cross-validation. In panels (b) and (c) comparison is also made to unpenalised maximum likelihood estimation. The horizontal lines in panels b and c are the large sample values. Note: Horizontal spread within each sample size grouping is just random jitter to aid display.

c

A screenshot of a cell phone

Description automatically generated

b

a

Table S1: Mean (sd) calibration slope and c-index in the validation data for models developed using ridge regression

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** | **Calibration Slope** | | | | | | | **c-index** | | | | | | |
| **Maximum Likelihood** | **5-fold CV** | **Repeat 5-fold CV** | **Bootstrap 5-fold** | **10-fold CV** | **Repeat 10-fold CV** | **Bootstrap 10-fold CV** | **Maximum Likelihood** | **5-fold CV** | **Repeat 5-fold CV** | **Bootstrap 5-fold** | **10-fold CV** | **Repeat 10-fold CV** | **Bootstrap 10-fold CV** |
| 100 | 0.260  (0.081) | 1.324  (0.786) | 0.713  (0.447) | 0.655  (0.231) | 1.261  (0.752) | 0.672  (0.416) | 0.620  (0.219) | 0.627 (0.026) | 0.648  (0.023) | 0.641  (0.025) | 0.642  (0.025) | 0.648  (0.023) | 0.641  (0.025) | 0.641  (0.025) |
| 200 | 0.488  (0.088) | 1.286  (0.533) | 0.904  (0.286) | 0.886  (0.205) | 1.212  (0.454) | 0.870  (0.272) | 0.852  (0.194) | 0.658  (0.018) | 0.669  (0.016) | 0.666  (0.017) | 0.667  (0.016) | 0.669  (0.016) | 0.666  (0.017) | 0.666  (0.017) |
| 300 | 0.610  (0.098) | 1.215  (0.380) | 0.992  (0.276) | 0.971  (0.205) | 1.164  (0.348) | 0.959  (0.262) | 0.939  (0.195) | 0.674  (0.014) | 0.681  (0.012) | 0.679  (0.012) | 0.679  (0.012) | 0.681  (0.012) | 0.679  (0.012) | 0.679  (0.012) |
| 400 | 0.678  (0.092) | 1.152  (0.259) | 0.999  (0.186) | 0.992  (0.170) | 1.109  (0.233) | 0.972  (0.179) | 0.965  (0.163) | 0.681  (0.010) | 0.686  (0.009) | 0.685  (0.009) | 0.685  (0.009) | 0.686  (0.009) | 0.685  (0.009) | 0.685  (0.009) |
| 500 | 0.728  (0.082) | 1.128  (0.222) | 1.014  (0.170) | 1.011  (0.145) | 1.093  (0.196) | 0.989  (0.163) | 0.987  (0.140) | 0.686  (0.009) | 0.690  (0.008) | 0.689  (0.008) | 0.689  (0.008) | 0.690  (0.008) | 0.689  (0.008) | 0.689  (0.008) |
| 600 | 0.763  (0.086) | 1.108  (0.188) | 1.020  (0.153) | 1.018  (0.144) | 1.076  (0.172) | 0.996  (0.146) | 0.996  (0.139) | 0.690  (0.006) | 0.693  (0.006) | 0.693  (0.006) | 0.693  (0.006) | 0.693  (0.006) | 0.693  (0.008) | 0.693  (0.007) |
| 700 | 0.788  (0.083) | 1.096  (0.179) | 1.028  (0.146) | 1.022  (0.134) | 1.068  (0.162) | 1.007  (0.140) | 1.001  (0.130) | 0.692  (0.006) | 0.695  (0.006) | 0.694  (0.006) | 0.694  (0.006) | 0.694  (0.006) | 0.694  (0.006) | 0.694  (0.006) |
| 800 | 0.803  (0.076) | 1.071  (0.148) | 1.020  (0.128) | 1.014  (0.118) | 1.045  (0.136) | 1.000  (0.123) | 0.995  (0.114) | 0.692  (0.006) | 0.696  (0.006) | 0.696  (0.006) | 0.696  (0.006) | 0.696  (0.006) | 0.696  (0.006) | 0.695  (0.006) |
| 900 | 0.828  (0.079) | 1.079  (0.143) | 1.031  (0.125) | 1.027  (0.118) | 1.055  (0.135) | 1.013  (0.120) | 1.009  (0.115) | 0.696  (0.005) | 0.697  (0.005) | 0.697  (0.005) | 0.697  (0.005) | 0.697  (0.005) | 0.697  (0.005) | 0.697  (0.005) |
| 1000 | 0.838  (0.074) | 1.063  (0.124) | 1.023  (0.113) | 1.021  (0.107) | 1.042  (0.119) | 1.006  (0.109) | 1.004  (0.104) | 0.698  (0.005) | 0.699  (0.004) | 0.699  (0.004) | 0.699  (0.004) | 0.699  (0.004) | 0.699  (0.004) | 0.699  (0.004) |

Table S2: Median (Median Absolute Deviation) calibration slope and c-index in the validation data for models developed using ridge regression

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** | **Calibration Slope** | | | | | | | **c-index** | | | | | | | |
| **Maximum Likelihood** | **5-fold CV** | **Repeat 5-fold CV** | **Bootstrap 5-fold** | **10-fold CV** | **Repeat 10-fold CV** | **Bootstrap 10-fold CV** | **Maximum Likelihood** | **5-fold CV** | **Repeat 5-fold CV** | **Bootstrap 5-fold** | **10-fold CV** | **Repeat 10-fold CV** | **Bootstrap 10-fold CV** |
| 100 | 0.255  (0.078) | 1.104  (0.556) | 0.612  (0.264) | 0.622  (0.213) | 1.026  (0.484) | 0.577  (0.264) | 0.589  (0.197) | 0.631  (0.024) | 0.652  (0.019) | 0.645  (0.022) | 0.645  (0.021) | 0.652  (0.020) | 0.644  (0.023) | 0.644  (0.021) |
| 200 | 0.485  (0.082) | 1.164  (0.363) | 0.844  (0.207) | 0.868  (0.178) | 1.112  (0.323) | 0.822  (0.208) | 0.835  (0.171) | 0.661  (0.017) | 0.671  (0.015) | 0.668  (0.016) | 0.668  (0.016) | 0.670  (0.015) | 0.667  (0.016) | 0.667  (0.016) |
| 300 | 0.599  (0.096) | 1.133  (0.290) | 0.930  (0.212) | 0.939  (0.186) | 1.083  (0.263) | 0.900  (0.208) | 0.909  (0.182) | 0.674  (0.013) | 0.683  (0.010) | 0.681  (0.011) | 0.681  (0.011) | 0.683  (0.010) | 0.681  (0.011) | 0.681  (0.011) |
| 400 | 0.669  (0.091) | 1.109  (0.232) | 0.990  (0.174) | 0.970  (0.169) | 1.070  (0.212) | 0.958  (0.167) | 0.941  (0.159) | 0.681  (0.010) | 0.688  (0.008) | 0.687  (0.008) | 0.687  (0.008) | 0.688  (0.008) | 0.687  (0.008) | 0.687  (0.008) |
| 500 | 0.718  (0.074) | 1.084  (0.169) | 0.982  (0.132) | 0.987  (0.126) | 1.061  (0.156) | 0.958  (0.123) | 0.961  (0.124) | 0.687  (0.009) | 0.691  (0.007) | 0.691  (0.007) | 0.691  (0.007) | 0.691  (0.007) | 0.690  (0.007) | 0.690  (0.007) |
| 600 | 0.761  (0.087) | 1.096  (0.187) | 1.008  (0.135) | 1.010  (0.136) | 1.071  (0.168) | 0.984  (0.130) | 0.990  (0.131) | 0.690  (0.007) | 0.694  (0.006) | 0.693  (0.006) | 0.693  (0.006) | 0.694  (0.006) | 0.693  (0.006) | 0.693  (0.007) |
| 700 | 0.781  (0.081) | 1.066  (0.150) | 1.010  (0.133) | 1.004  (0.127) | 1.043  (0.143) | 0.991  (0.130) | 0.988  (0.124) | 0.693  (0.006) | 0.695  (0.006) | 0.695  (0.006) | 0.695  (0.006) | 0.695  (0.006) | 0.695  (0.006) | 0.695  (0.006) |
| 800 | 0.793  (0.068) | 1.045  (0.124) | 1.010  (0.112) | 0.998  (0.104) | 1.026  (0.117) | 0.991  (0.109) | 0.979  (0.099) | 0.695  (0.006) | 0.696  (0.005) | 0.696  (0.005) | 0.696  (0.005) | 0.696  (0.005) | 0.696  (0.005) | 0.696  (0.005) |
| 900 | 0.825  (0.078) | 1.065  (0.130) | 1.019  (0.118) | 1.018  (0.117) | 1.043  (0.125) | 1.001  (0.115) | 1.000  (0.114) | 0.697  (0.005) | 0.698  (0.004) | 0.698  (0.004) | 0.698  (0.004) | 0.698  (0.004) | 0.698  (0.004) | 0.698  (0.004) |
| 1000 | 0.838  (0.072) | 1.059  (0.121) | 1.019  (0.103) | 1.019  (0.101) | 1.034  (0.118) | 1.001  (0.100) | 1.001  (0.096) | 0.698  (0.005) | 0.699  (0.004) | 0.699  (0.004) | 0.699  (0.004) | 0.699  (0.004) | 0.699  (0.004) | 0.699  (0.004) |

Table S3: Median Absolute Deviation in Lambda for models developed using ridge regression

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **N** | **5-fold cross-validation** | | | **10-fold cross-validation** | | |
| **Single CV** | **50-Repeat CV** | **Bootstrap CV** | **Single CV** | **50-repeated CV** | **Bootstrap** |
| 100 | 0.2299 | 0.0820 | 0.0541 | 0.1951 | 0.0706 | 0.0488 |
| 200 | 0.1053 | 0.0501 | 0.0336 | 0.0913 | 0.0454 | 0.0300 |
| 300 | 0.0597 | 0.0371 | 0.0244 | 0.0514 | 0.0339 | 0.0228 |
| 400 | 0.0420 | 0.0279 | 0.0191 | 0.0341 | 0.0252 | 0.0170 |
| 500 | 0.0305 | 0.0215 | 0.0129 | 0.0241 | 0.0190 | 0.0118 |
| 600 | 0.0281 | 0.0182 | 0.0121 | 0.0206 | 0.0166 | 0.0107 |
| 700 | 0.0197 | 0.0157 | 0.0109 | 0.0159 | 0.0147 | 0.0097 |
| 800 | 0.0167 | 0.0138 | 0.0082 | 0.0123 | 0.0114 | 0.0073 |
| 900 | 0.0146 | 0.0116 | 0.0080 | 0.0130 | 0.0103 | 0.0072 |
| 1000 | 0.0143 | 0.0104 | 0.0075 | 0.0116 | 0.0094 | 0.0056 |